A Short Note on Surveillance Testing

FD, CG, VK & NQ on behalf of LMH Governing Body

16/09/2020

Introduction

Controlling the spread of coronavirus infection within the university requires a reliable means of detecting clusters of infection. Guiding future policy changes, e.g. restricting in-person instruction, demands accurate information about the prevalence of infection. This document outlines some simple strategies to achieve both goals. We focus on detecting the presence of infection among *household groups* of students in college or university-provided residential accommodation. We restrict attention to students rather than the university community as a whole because, given limited testing and logistical capacity, it makes sense to focus on the sub-population in which transmission is most likely to occur. We treat household groups, rather than individual students, as our unit of analysis for two reasons: first because this allows us to employ a simple and efficient pooled testing strategy, and second because it is transmission *between* rather than within households that is of more serious concern. The problem of estimating infection prevalence among households and detecting clusters of infection is fundamentally different from that of determining whether a particular individual has the coronavirus. In particular, the high test specificity and sensitivity along with rapid turnaround that are crucial for individual medical diagnosis are not absolutely essential for surveillance testing. Our calculations suggest that relatively modest resources may be sufficient to obtain valuable information, separate from and parallel to the "Early Alert Testing" service. The ideas we discuss are neither novel nor original, and we do not propose a detailed plan for surveillance testing. Our goals are primarily to explain randomized and pooled testing strategies to a non-specialist audience, and to calculate the approximate scale of testing required. We begin by abstracting away the problem of false negatives and false positives. The final section discusses how these affect our analysis.

Bottom Line Recommendations

We view a policy of universal pooled testing, like that planned by the University of Cambridge, as the first-best policy for identifying clusters of infection and understanding changes in coronavirus prevalence among our students. Sufficient laboratory capacity for roughly 1500 tests per week would allow fortnightly universal testing of students in residential accommodation, pooling tests by household group. Because capacity cannot be diverted from

the Early Alert Testing service, the university would need to obtain around 12,000 tests from an alternative source over the course of MT. These need not have a rapid turnaround time: a delay of several days would still allow us to assess the spread of infection. If insufficient laboratory capacity is available for a policy of universal testing, randomized testing on the scale of around 300 tests per week (2400 over the course of the term) would provide valuable information about changes in prevalence and, combined with an adaptive sampling scheme, help us to identify clusters of infection as they emerge.

Tests should be pooled by household.

Pooled testing is a procedure that can be helpful in situations where laboratory capacity, rather than the ability to collect test samples, is the limiting factor in expanding testing. In this procedure, a single laboratory test is carried out on the *combined* samples of a group of individuals. For a sufficiently accurate test, a negative result for the pooled sample indicates that no one in the group is infected while a positive result indicates that at least one person in the group is infected. Traditionally, a positive group test result would be followed by further testing to determine which individual or individuals in the group is infected. If, however, we are only interested in detecting infected *groups*, no further testing is required. This makes pooled testing an extremely efficient procedure when applied to household groups: only one test is required to determine whether any members of a given household are infected with coronavirus. Given that household members would be required to isolate if any of them tested positive, for the purposes of reducing the spread of infection it is irrelevant precisely which of them is infected.

There are approximately 24,000 students at the University of Oxford divided roughly evenly between undergraduate and graduate students.[1] As a rough approximation, suppose that 100% of undergraduates and roughly 50% of graduate students are housed in college or university–provided accommodation. With a household size of 6, this makes 3000 households in total. Accordingly, sufficient laboratory capacity for 1500 tests per week would allow fortnightly universal surveillance testing, pooling tests by household. Pooled testing reduces the number of laboratory tests required but does not affect the number of samples that need to be collected. To reduce the burden of collecting samples from approximately 200 students daily, swabs could be self-administered under the guidance of a healthcare professional.[2]

We view universal pooled testing of households as the first-best policy for understanding and controlling the spread of infection. Should there be insufficient testing capacity to implement such a policy, however, there are alternative strategies that can provide valuable information with fewer resources. The remainder of this document explains some simple and effective possibilities.

[1] For precise student numbers, see https://www.ox.ac.uk/about/facts-and-figures/student-numbers.

[2] This is the procedure used by the ONS in their Coronavirus Infection Survey pilot. A recent study suggests that this procedure is no less accurate than having healthcare workers conduct the swabs directly: https://www.medrxiv.org/content/10.1101/2020.04.11.20062372v1.

Absent universal testing, households could be randomly sampled.

Suppose that C of the roughly 3000 households in the university contain at least one member who is infected with coronavirus. With a perfect test, we could learn C exactly with 3000 tests, pooling by household. But what if we only have the capacity to carry out n tests, where n is much smaller than 3000? In this case we cannot learn C exactly, but by employing *randomized testing* we can produce an estimate that is sufficiently accurate to guide our mitigation policies and help us to identify clusters of infection. The value of randomized testing is widely acknowledged. It underlies both the Coronavirus Infection Survey pilot conducted by the ONS, and the REACT-1 survey conducted by the Department of Health and Social Care. Both of these surveys are extremely valuable, but as they aim to be nationally representative, neither employs a sufficiently large sample size to track local trends. As such we propose that randomized surveillance testing be carried out within the university.

The most basic form of randomized testing relies on a *simple random sample*. To use this procedure, we assign each household an id number, place all 3000 id numbers into a hat, mix them up thoroughly, and then draw out n at random. Each of these n households is sent for a pooled coronavirus test. We then use the number positive tests results in our sample to construct an estimate E of the number of infected households in the population. For example, if 1 household in a sample of 300 tests positive, then we would estimate that 10 households out of 3000 in the university have coronavirus. Repeating this process weekly would allow us to estimate how the prevalence of coronavirus changes over time.

Because households are chosen at random, the estimated number of infected households E may not equal the true number of infected households C. Purely by chance, we might draw a sample of households in which no one is infected. In this case E would equal zero: an underestimate of C. We could also be unlucky in the opposite direction and, purely by chance, draw a sample that contains everyone at Oxford who is infected. In this case E would be an overestimate of C. Both of these possibilities, however, are remote: E is highly likely to be close to C. With an appropriate sample size, an estimate based on random testing is reliable enough to tell us with high confidence whether C is, say, greater than 35. It can also be used to determine whether prevalence is increasing and, if so, how quickly.

Testing only those with symptoms will not suffice.

Relative to the Early Alert Testing system, which tests only symptomatic individuals, a crucial advantage of randomized testing, or universal testing, is that it will also detect pre-symptomatic and asymptomatic cases. This is important because, as explained in the SAGE report of 3rd September, "asymptomatic transmission is a key risk in university settings." Indeed, a recent study cited in the SAGE report finds that only 18% of those aged 0–19 and 22% of those aged 20–39 who are infected with coronavirus show symptoms of the disease.[3] Even if these figures are substantial overestimates, it is likely that a considerable fraction of infections among our students will go undetected by the Early Alert Testing system. While

[3] https://arxiv.org/abs/2006.08471

most students themselves are at low risk of serious complications from coronavirus, members of staff are at substantially higher risk. By the time the Early Alert system registers an increase in cases among members of staff, it may already be too late for mitigation policies, such as curtailing in-person instruction, to be effective. Even if their turn-around times were somewhat slower than those for tests conducted through the Early Alert system, randomized pooled tests could still provide more timely information about the spread of infection, given the likely extent of asymptomatic transmission among our students.

Roughly 300 tests per week would provide useful information.

The key question in designing a randomized testing protocol is how large a sample size to use. To answer this question, we first need a rough idea of the prevalence of coronavirus among households at the beginning of term. Based on recent results from REACT-1, approximately 0.25% of individuals in the 18-24 age group had the coronavirus at the start of September. Taking this as a lower bound for prevalence in early October, suppose that at least 45 of the roughly 18,000 students in residential accommodation has coronavirus at the start of MT. Given that our interest is in households rather than individuals, the question remains: if 45 students have coronavirus, then how many *households* are infected?

The answer depends on what we are willing to assume about households. At one extreme, as many as 45 households could be infected: this would require that exactly *one* person in each affected household has the coronavirus. At the other extreme, it is possible that as few as 8 households could be infected. This would require 7 households in which every member was infected, and 1 household with 3 infected members. During term, intra-household transmission will likely be substantial.[4] At the beginning of term, however, it is unlikely that cases will cluster strongly within households, as most groups will not have met in person over the summer. If initial coronavirus cases can be viewed as independently and uniformly assigned to households, then 45 infected students would translate into approximately 44.7 infected households on average.[5] To allow for some initial clustering of cases, suppose that there are initially 35 infected households, yielding a household prevalence of approximately 1.16% at the start of term.[6]

Given a rough idea of household prevalence at the start of MT, there are several simple ways to approximate the sample size required for randomized pooled testing to provide useful information. The first and simplest is to ask how many households we would need to test, on average, before obtaining our first positive result. If C out of 3000 households are infected at the start of term, then the answer is approximately $3000/C$.[7] Table 1 provides exact results over a range of values for C. This calculation gives a rough-and-ready method for showing

[4] A secondary infection rate of 38% within households would be consistent with a recent study of patients in New York State: https://doi.org/10.1093/cid/ciaa549.

[5] Specifically, if N is the number of households, k is household size, and S is the number of infected students, the expected number of infected households is $N\left[1-\binom{kN-S}{k}/\binom{kN}{k}\right]$.

[6] While we consider an initial figure of 45 more likely that 35, using the smaller value builds an extra margin of conservatism into our calculations: larger samples are needed to detect cases at a lower prevalence.

[7] The exact value is $(N+1)/(C+1)$ for a population of size N containing C positives.

that a proposed sample size is likely *too small* to be useful. For example, a sample size of 50 households would be too small to provide useful information when C is close to 35 because we would, on average, require approximately 83 samples to obtain our first positive result if this were the true number of infected households.

Table 1: Expected number of samples before the first positive result under random sampling without replacement from a population of size 3000.

Infected Households (C)	Prevalence (%)	Expected #Samples
25	0.83	115
30	1.00	97
35	1.17	83
40	1.33	73
45	1.50	65
50	1.67	59

A more refined way to determine an appropriate sample size is by calculating *how close* the estimated number of infected households, E, will likely be to the true number of infected households, C, for a given sample size. Suppose, as before, that there are initially 35 infected households. With a sample size of $n = 300$ the estimated number of infected households will fall in the range $[20, 60]$, inclusive, with approximately 82% probability. Moreover, with this sample size there would be only a 2% chance of *failing* to detect any infected households.[8]

A third way of deciding how many samples are needed is by calculating the chance that we would be able to reliably distinguish between two different levels of coronavirus prevalence. Suppose that we started term with 35 infected households. If this number doubled to 70, with a sample size of $n = 300$ we would have approximately a 72% of detecting an increase, with high confidence. If the number of infected households trebled to 105, our chance of detecting an increase with 300 samples would rise to 96%.[9]

Broadly speaking, our calculations suggest that a simple random sample of approximately 300 household groups per week (2400 over MT) would provide a considerable amount of useful information. This would amount to 10% of the number of tests required for weekly universal testing and 20% of the number required for fortnightly universal testing.

[8]The results in this paragraph are calculated from the quantiles of a Hypergeometric(N, C, n) distribution where N is the population size, C is the number of positives in the population, and n is the sample size.

[9]These values correspond to a power calculation based on Fisher's exact test for a Hypergeometric(N, C, n) distribution where N is the population size, n is the sample size, and C is the number of positives in the population. For $N = 3000$ and $n = 300$ a test of $H_0 \colon C \leq 35$ versus $H_1 \colon C > 35$ with a critical value of 5 has size $\alpha = 0.13$. (Due to discreteness it is impossible to achieve a size of precisely 0.1). The power of this test is 0.72 against the alternative $C = 70$ and 0.96 against the alternative $C = 105$.

An adaptive sampling scheme could identify clusters of infection.

Our discussions in the preceding section focused mainly on estimating the *prevalence* of infection among households. Another, and arguably more important, goal is to identify clusters of infection as they emerge in the university. As it turns out, both of these aims can be achieved simultaneously by using a more sophisticated sampling procedure.

Given what we know about the spread of coronavirus, it is likely that infected households will begin to *cluster* within accommodation blocks and possibly colleges as the term progresses. Suppose that we test a simple random sample of 300 households during week 1 of MT. If there are roughly 35 infected households at the start of term, then we have a 98% of detecting *at least one of them* in this first round of testing. Let P be the households that test positive. In week two, rather than using all 300 of our weekly tests for a second round of random sampling, we could divert some of them to systematically testing the *neighbors* of P: say the households lodged on adjacent floors of the same building. If any of these neighboring households test positive, we could then go on to test *their* neighbors, and so on. This idea is called *adaptive cluster sampling.*[10]

For a fixed number of tests each week, incorporating adaptive sampling based on the previous week's results would increase our likelihood of identifying infected households. Moreover, it would still leave capacity for randomized surveillance testing to estimate changes in prevalence.[11] If, for example, we detected 4 infected households in week one and each household has 4 neighbors, we would still have 284 tests available for randomly sampling households in week 2.

False positives and negatives do not invalidate our analysis.

Thus far we have abstracted away the problem of false positives and false negatives. As our primary goal is to estimate prevalence and identify clusters of infection rather than to diagnose particular individuals, this is less of a concern. First, if false positive and false negative rates are constant, we can still gain useful information by examining *changes* in estimated prevalence. Second, recent research has proposed estimates of these rates that can be used to correct estimates of population prevalence. Third, false positives are not especially likely under pooled testing of households. In the appendix, we provide some back-of-the envelope calculations in support of this claim.

[10] Thompson, S.K. (1990), "Adaptive Cluster Sampling", *Journal of the American Statistical Association*, 85(412), pp. 1050-1059.

[11] Although the procedure is somewhat involved, it is possible to use results for *all tested households*–both those sampled randomly and those sampled in response to a positive test among their neighbors–to inform estimates of prevalence: see Thompson (1990).

Appendix: Approximating False Positive Rates

The likelihood of a false positive depends on three factors: the *sensitivity* and *specificity* of a test along with the prevalence of coronavirus in the relevant population. Defining

$$B = \frac{(100\% - \text{Specificity}) \times (100\% - \text{Prevalence})}{(\text{Sensitivity} \times \text{Prevalence})},$$

the false positive rate is given by $[B/(1+B)] \times 100\%$. Sensitivity is defined as the share of people who test *positive* among those who *have coronavirus*. Estimates of the sensitivity of PCR tests vary, but range between 71% and 98%.[12] Specificity is the share of people who test *negative* among those who *do not have coronavirus*. We can bound the specificity of PCR tests fairly accurately using results from the ONS Infection survey pilot. Between 1 June and 12 July, only 50 of the 122,776 samples collected as part of this survey tested positive. To compute a lower bound for the specificity, suppose that all 50 positive results were false positives. If so, this would mean that the ONS obtained 122,726 negative test results in a sample of 122,776 people who did not in fact have coronavirus, implying a specificity of 99.96%. If any of the positives were true positives, then the true specificity must be higher.

The figures in the preceding paragraph are for "single test" errors, i.e. errors that arise when testing individuals. In contrast, our proposals in this document involve *pooling* samples from household groups. The question then becomes: how do single test errors translate into pooled test errors? One important concern is dilution errors, the possibility that pooling samples could lower sensitivity by reducing the concentration of material from each individual sample. For modest pool sizes such as those proposed above, however, recent studies find no evidence of reduced sensitivity.[13] Unless cross-contamination of group samples is more common than cross-contamination of individual samples, pooling should presumably have no effect on specificity. Indeed, if we were to adopt a policy of re-testing the individuals in pools that test positive, sensitivity would *increase* relative to individual testing, because it is extremely unlikely to obtain more than one false positive in a row.[14]

Based on the arguments of given above, assume a worst-case sensitivity of 71% for pooled testing and a worst-case specificity of 99.96%. As in our discussions in the body of the document, suppose there are 35 infected households out of 3000 at the start of term, for an initial prevalence of roughly 1.16%. Under these assumptions, the household false positive rate would be 4.6%. In other words, our *worst case* estimate is that 1 out of 20 households instructed to isolate will have been needlessly inconvenienced under a policy of testing asymptomatic households. This does not strike us as a particularly worrying prospect.

[12] https://www.bmj.com/content/369/bmj.m1808
[13] See Pikovski & Bentele (2020): https://doi.org/10.1017/S0950268820001752.
[14] Ibid.